Análise de dados com suporte limitado: modelos *power logit* e contribuições à inferência robusta

Francisco Felipe de Queiroz

Tese Apresentada
ao
Instituto de Matemática e Estatística
da
Universidade de São Paulo
para
Obtenção do Título
de
Doutor em Ciências

Área de Concentração: Estatística

Orientadora: Profª. Drª. Silvia Lopes de Paula Ferrari

Durante o desenvolvimento deste trabalho o autor recebeu auxílio financeiro da CAPES e do CNPq

São Paulo, Julho de 2022

Análise de dados com suporte limitado: modelos *power logit* e contribuições à inferência robusta

Versão original simplificada

Esta é a versão original da tese elaborada pelo candidato Francisco Felipe de Queiroz, tal como submetida à Comissão Julgadora.

Dedico este trabalho à minha avó, Antônia,
e à minha mãe, Lúcia, pelos
ensinamentos e valores que me foram passados.

Agradecimentos

À minha avó Antônia, minha mãe Lúcia, minha madrinha Nenzinha e meu tio João, pelo apoio, amor e por sempre acreditarem em mim. Essa conquista é nossa!

À minha orientadora, Silvia Ferrari, pela orientação impecável e segura, confiança, amizade, disponilidade e dedicação. Obrigado pelos conselhos e ensinamentos compartilhados durante o doutorado. Tenho muito orgulho de lhe ter tido como orientadora!

Aos professores do Departamento de Estatística da UFRN, em especial a Dione Valença, Artur Lemonte, André Pinho, Pledson Guedes, Carla Vivacqua, Damião Nóbrega, Iloneide Ramos, Luz Milena, Bruno Monte e Moisés Medeiros. Agradeço pela amizade, disponibilidade em ajudar e pelos conhecimentos compartilhados durante minha graduação e mestrado em Estatística. Tenho orgulho de ter sido aluno de vocês.

Aos professores do Departamento de Estatística do IME-USP, em especial aos professores Alexandre Patriota, Silvia Ferrari, Gilberto de Paula e Júlia Pavan pela excelência no ensino e pelos conhecimentos compartilhados durante os cursos do Doutorado.

À Regiane, pela simpatia e eficiência com assuntos da pós-graduação do IME-USP.

À segunda família que encontrei em São Paulo, Giovanni, Rosária e Gabriella. Vocês tornaram essa caminhada menos difícil. Em especial a Giovanni, pelo companheirismo e amor que me tem dedicado.

À minha querida amiga Joelma, pelo amor, carinho e irmandade. Ao meu pequeno afilhado, Nicolas, pelos momentos de alegria que me proporciona. À Maisa, pelo carinho e por se fazer sempre presente. À Flávia, Inara, Joyce e Vanessa, pela amizade sincera. Aos amigos Carla, Lucas, Erika e Daniel pelo companheirismo.

Aos amigos que fiz durante o Doutorado em São Paulo, Luisa Borsato, Rodrigo Garcia, Cátia Tondolo, Gabriela Vasconcelos, Magno Tairone e Luiza Tuller. Em especial, à Luisa Borsato pela amizade sincera, disponibidade em ajudar e suporte. Essa caminhada seria bem mais difícil sem vocês.

À minha professora de conversação em inglês do Cambly, Leah, pelas aulas durante quase três anos do doutorado, por me ajudar a melhorar o inglês e me sentir mais confiante.

À Capes e ao CNPq pelo auxílio financeiro.

Resumo

Queiroz, F. F. **Análise de dados com suporte limitado: modelos *power logit* e contribuições à inferência robusta**. 2022. 191f. Tese (Doutorado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2022.

Dados contínuos limitados, particularmente no intervalo unitário, aparecem em diferentes áreas, incluindo ecologia, biologia, economia e saúde pública. Alguns exemplos são a fração da cobertura vegetal, a proporção da renda familiar gasta em planos de saúde e a prevalência de doenças crônicas. Estes dados geralmente são altamente assimétricos, possuem dispersão dependendo da média e muitas vezes apresentam valores nas fronteiras. Modelos de regressão que utilizam a distribuição beta são amplamente empregados em aplicações. A regressão beta permite a interpretação direta dos parâmetros, acomoda assimetria e heterocedasticidade, sendo razoavelmente flexível. A inferência em modelos de regressão beta geralmente é baseada em métodos de máxima verossimilhança ou Bayesianos, para os quais a informação dos dados vem da função de verossimilhança. Em ambos os casos, a inferência pode ser altamente influenciada por observações atípicas. O procedimento de inferência pode então ser substituído por um método robusto ou pode-se empregar modelos baseados em distribuições mais flexíveis do que a distribuição beta. Nesta tese, contribuímos para a modelagem estatística de dados limitados em duas direções. Primeiramente, definimos e estudamos os modelos *power logit*, uma classe altamente flexível de modelos de regressão com parâmetros interpretáveis adequados para modelagem de dados limitados com diferentes características. São apresentadas medidas de diagnóstico e de influência, e um novo pacote computacional é desenvolvido. Apresentamos também os modelos de regressão *power logit* inflacionados, que podem ser empregados quando os dados incluem observações em um dos extremos do suporte. A segunda parte desta tese é dedicada ao desenvolvimento de métodos inferenciais robustos em regressão beta inflacionada. Os estimadores propostos possuem boas propriedades e apresentaram bom desempenho em experimentos de simulação. Rotinas computacionais para uso dos estimadores propostos são fornecidas.

Palavras-chave: Dados fracionários, Inferência robusta, Proporções contínuas, Regressão beta, Regressão beta inflacionada.

Abstract

Queiroz, F. F. **Bounded continuous data: power logit models and contributions to robust inference**. 2022. 191f. Tese (Doutorado) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2022.

Bounded continuous data, particularly on the unit interval, appear in different areas, including ecology, biology, economics, and public health. Some examples are the fraction of vegetation cover, the proportion of family income spent on health plans, and the prevalence of chronic illness. The data are usually highly skewed, have dispersion depending on the mean, and often present values at the boundaries. Regression models that use the beta distribution are widely employed in applications. Beta regression allows direct parameter interpretation, asymmetry and heteroscedasticity while reasonably flexible. Inference in beta regression models is usually based on maximum likelihood or Bayesian methods, for which the information from the data comes from the likelihood function. In either case, the inference can be highly influenced by atypical observations. The inference procedure may then be replaced by a robust method, or one may employ models based on more flexible distributions than the beta distribution. In this dissertation, we contribute to the statistical modeling of bounded data in two directions. First, we define and study the power logit models, a highly flexible class of regression models with interpretable parameters suitable for modeling bounded data with different characteristics. Diagnostic and influence measures are presented, and a new computational package is developed. We also present the inflated power logit regression models, which may be employed when the data include observations at one of the extremes of the support set. The second part of this dissertation is devoted to developing robust inference methods in inflated beta regression. The proposed estimators have good properties and performed well in simulation experiments. Computational routines for using the proposed estimators are provided.

Keywords: Beta regression, Continuous proportion, Fractional data, Inflated beta regression, Robust inference.